FINAL PROJECT

Antonio Cadavid

Alejandro Suarez

Jairo E. Rodríguez

Juan D. Hernandez

A00354484

A00359653

A00354217

A00356210

CAPSTONE PROJECT

ICESI UNIVERSITY

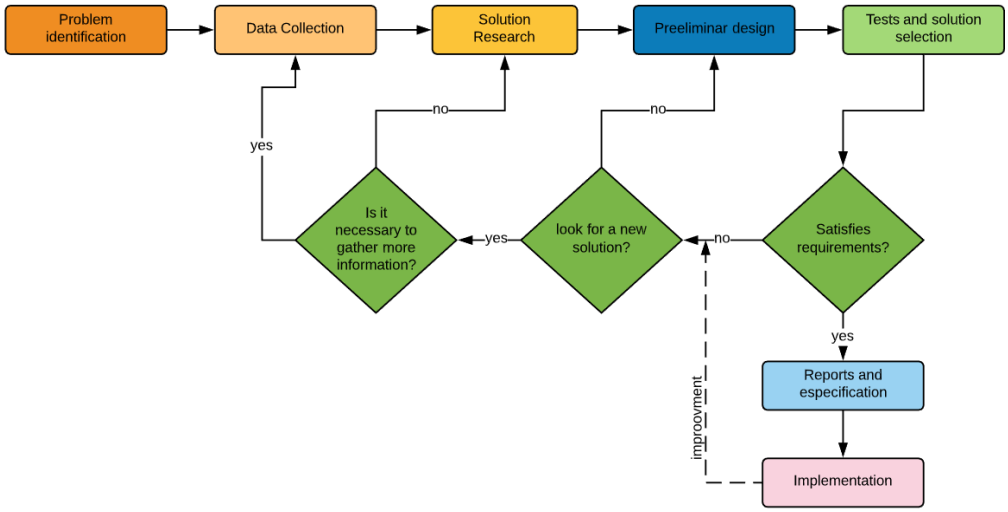
Bank Marketing

Bank Marketing

Problematic context

To create a solution to the problem we will tackle we chose to use the Engineering method to develop following a systematic focus, according to the problem situation.

Based on the description of the engineering method from the book “Introduction to Engineering” from Paul Wright, the following flow diagram has been defined, which steps we will be following during the development of our solution.



Step 1: Identification of the problem

Problem definition:

The CEO of Wekebank inc has noticed that one product in their portfolio is good profit wise and could increase a lot the banks wealth, this product is the famous “term deposit”. First, the CEO wants to be able to visualize the attributes of the client database in a graphic way so the bank's board can have a better understanding of them. Since the term deposit will generate a lot of revenue for the bank, the CEO needs a model that can determine if a client will or will not acquire a subscription to a term deposit based on his/her attributes. The goal is to spot these clients so the marketing team of the bank can focus on them and develop an aggressive marketing campaign to sell as many term deposits subscriptions as possible.

For this reason, the CEO has hired you and your team of engineers to develop a solution to tackle this problem and help the bank to increase its revenue by selling more subscriptions to term deposits.

Problem identification:

The Bank needs to increase its profit and take advantage of their client´s database.

Also, it needs to visualize the data of their clients in a series of charts.

Identification of the needs:

-The bank needs to visualize the data in a series of charts and graphics.

-The bank needs to filter the information from a desired way.

-The bank needs to determine whether a client will or will not acquire a subscription to a term deposit.

Step 2: Data collection:

In this step we will compile some information through research to understand the problem better and develop a better solution.

For us as an engineering team is important to deliver the best possible solution for this problem, it is a need to understand the product which sales they are trying to increase, also the classification method that we are going to use in order to infer the decision of the clients, in fact this is one, if not the most crucial point in the whole development since we

must develop a high accuracy prediction model, this is critical. Since we need to classify variables and infer their classes based on parameters, we find a decision tree really suitable to develop an optimal solution for this problem due its strengths. Decision trees are able to perform classifications without requiring a lot of computation due its recursive nature, that is an advantage due the large amount of records that we will handle, also they are able to generate clear and understandable classification rules, which is pretty important since this solution is implemented and maintained by humans.

What Is a Term Deposit?

A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.

The investor must understand when buying a term deposit that they can withdraw their funds only after the term ends. In some cases, the account holder may allow the investor early termination—or withdrawal—if they give several days notification. Also, there will be a penalty assessed for early termination.

Examples of term deposits include certificates of deposit (CDs) and time deposits.

- A term deposit is a type of deposit account held at a financial institution where money is locked up for some set period.
- Term deposits are usually short-term deposits with maturities ranging from one month to a few years.
- Typically, term deposits offer higher interest rates than traditional liquid savings accounts, whereby customers can withdraw their money at any time. (Chen, 2020)

For the sake of the solution it was necessary for us to understand a little bit of what a term deposit really is, that's why we researched a bit about this financial product. We just find it more comfortable to develop the solution knowing what the bank is really trying to sell, it gives us a more global approach to the understanding of the problem.

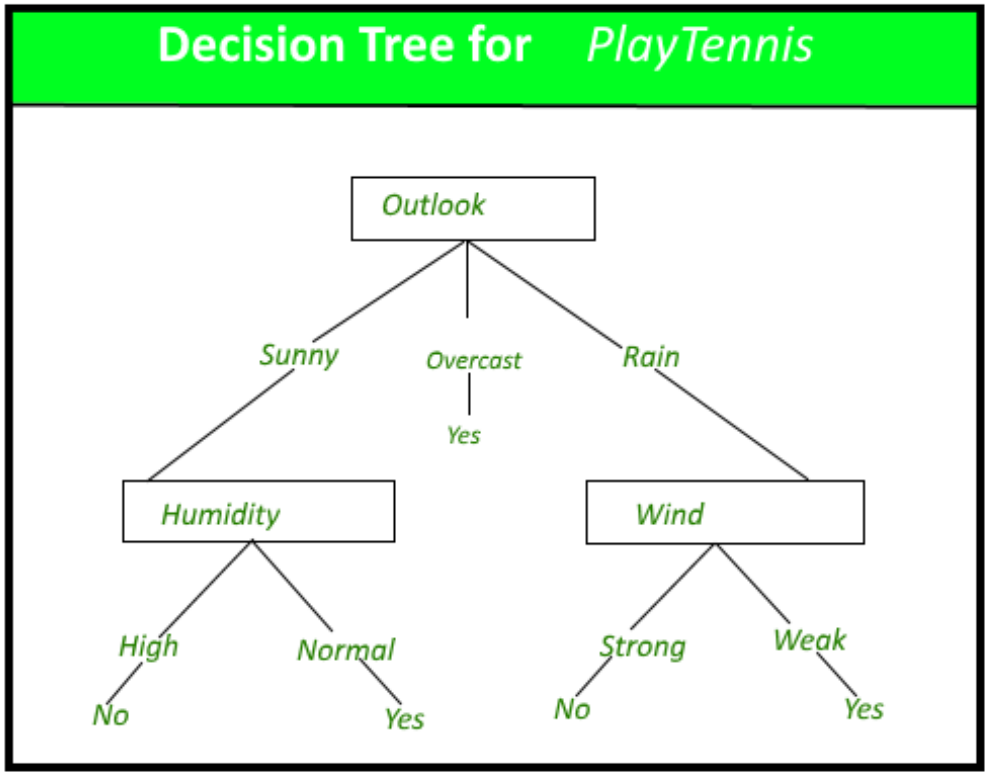
Decision tree

Decision Tree: Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. (Geeks for geeks, 2019)

Construction of Decision Tree:

A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification. (Geeks for geeks, 2019)

Decision tree representation:



(Geeks for geeks, 2019)

State of the art:

Since we chose to implement our own version of a decision tree to classify the variables we researched to see if there were hardcore tested libraries that have an implementation of a decision tree in order to grasp a better understanding of the data structures and algorithms

used in the implementation, so our own is a proper one. So here are ones we found and checked.

Decision Tree 3.4.3 is a Python module for decision-tree based classification of multidimensional data that we found. It is generic but some of the pros is that´s been heavily tested. And that is good if our code would be written in Python but ours is written in C#, regardless of that we can study its components and its implementation, taking into account that Python is a lighter language that is optimized for this type of tasks.

Accord .NET is a .NET machine learning framework combined with audio and image processing libraries completely written in C#. This is a pretty professional piece of software that has been used in a wide number of projects, it certainly gave us an insight of what we should do to implement ours.

Machine learning is a field that gets bigger and bigger each day. There is plenty of code out there that has been written to tackle issues and problems in this field. Getting to research the work of other people in this field gave us better insights on the same and helped us to feel mor confident to implement our own

Step 3: Solution research

We will tackle the problem of CEO of Wekebank using different approaches.

Alternative 1: In this approach, we will start with an aesthetically pleasing graphic interface for the user where we will ask that select from his directory the database to load the dataset of the company. The information is categorized by columns with: Age, job, marital, education, debit, balance, housing, loan, deposit. The dataset will be displayed by columns and also will allow you to filter by any category that you need: Ex: Category: Age, Data: 20. After that the dataset only shows the data that we ask. Also, the graphic interface has two computer tabs that you can select if you need the dataset or the graphic with the global summary with all of data.

Alternative 2: In this approach, we will start with a simple interface where we will load the default dataset, it will not load irrelevant information about the problem to reduce the memory use. Also, we will generate the respective graphics to give simple summary about all the information. The program can filter by category already defined, but if it is a numerical variable it will display the values from the lower to the highest, on the contrary will be lexicographic.

Alternative 3: In this approach, we will start with a complex interface where we will try to show all the information from the dataset, but what makes it different is the way to filter. In the other alternatives, the tool to filter is a Combo Box that have the category to filter, in this case won't be like that, because the program allows to filter only with a click in the column of the datagrid. If you click once the data will organize by ascendingly, but if you

click it twice dancingly. Also, the program gives you a summary with a simple graphic to try to remember better the information.

Alternative 4: In this approach, we will start with an interactive interface where the user provides all the information that it needs. Like other alternatives, the user searches the dataset that he likes to show, after that the user goes through a checkbox with the category, the program allows to select the category that the user would like to see about the information, e.g. the sex, the salary or the age of an individual. We think that we have more control of all data and regulate the order. Also have the summary with different types of graphics that you need to see.

Alternative 5: In this approach, we will start with a simple interface where the person already loads the Dataset on the program. First the program allows the user to organize the information by a specific condition like: Age, birth, etc. and also allows to filter again if need it´s needed. The special of this alternative is in the summary, when you have all the information loads in the program you can select a special graphic to show like a: Circular, bar, waves, etc. and allows to select the information that the user needs graphically, and if needed allows to select the data separate or together. The program finish with a simple conclusion about all the generated graphics.

Step 4: Preliminary design

In this step, we will seek to analyze and evaluate which alternative is more feasible to carry out for the optimal development of the program.

Alternative 1. This alternative is optimal for the proper development of the application, but it has a drawback. The dataset may not be executed correctly, when trying to give the client the freedom to search for the dataset, there may be an error when downloading or cloning the repository where the program is located with all its files (including this one) and due to some carelessness it may be deleted without realizing it and the program would no longer be of any use at all, since the dataset is essential for the execution of the program.

Alternative 2: This alternative is really very good, since it allows us to save the user the search for the correct dataset for the correct execution of the program, it also has a graphical interface that is pleasing to the user and does what corresponds to the functional requirements of the program like: Filter, graph and others.

Alternative 3: This alternative is not good for the program, since it inhibits us from necessary functionalities, to say that it becomes a bit complex when it comes to interpreting how things are done. The way of filtering is not adequate, since it simply orders ascending and descending, that in quantitative variables, but in qualitative it orders lexicographically no more.

Alternative 4: This alternative gives the user total control of what he really wants to see and this can affect the interpretation that he wants to give regarding the Dataset, the user controlling what he really wants to see can cause that at the time of predicting Which user would or would not buy the product through the decision tree, we can find an error when

executing it, since without one of its variables it can affect all the learning that is implemented in the tree and can give a totally unexpected decision.

Alternative 5: In this alternative, it seems to me a very good idea to implement in the future, the way to give it the privilege of wanting to graph the information as you want and what data you want to visualize, so it does not affect the correct reading of the dataset and also provides better functionalities when it comes to graphing the results.

The alternatives that we are going to discard will be 1 and 3. This is because they are not optimal and have many defects, so implementing them would be a waste of time.

Step 5: Solution Selection

Now, we are going to qualify each alternative by some criteria to choose the best. The criteria have been chosen thinking about the user's experience with the program and how automated it can be.

- Criterion #1: Completeness of the program.
    1. Less than 50% of the requirements.
    2. Between 50% and 70% of the requirements.
    3. More than 70% of the requirements.

- Criterion #2: Complexity for the user.
    1. Very complex, the user can be confused by the number of elements in the interface.
    2. Complex, the interface is intuitive, but the user may not know how to handle the program at the first time.
    3. Nothing complex, any user could use it.
- Criterion #3: Capacity of the program to update itself with the database.
    1. Static, the database cannot be updated.
    2. The database can be updated but the program will not have a good ability to update.
    3. The database can be updated, and the program will have the ability to train with the new data

Alternative	Criterion #1	Criterion #2	Criterion #3	Total
2	3	3	3	9
4	3	3	2	8
5	3	2	2	7

In conclusion, we will choose alternative #2 since it has had the best score in the evaluation.

Reflective synthesis:

The statement of the integrating project for this semester was similar to choosing a data set (multivariate dataset) based on the fact that a problem could be defined and also solved with topics learned during the course.

This project consisted of an information analysis problem, specifically it had to deal with classification. To do this, the solution had to be implemented using the decision trees technique, in which the group, made up of 4 people, had to work to build their own decision tree and implement another with a C # library. On the other hand, I had to carry out experiments that would allow me to compare results obtained under the results obtained by the two classification modules mentioned above. These had to have 4 phases: Design of experiment, execution of the experiment, analysis of the results obtained and finally the evaluation.

Regarding the documentation to be delivered, this project required an engineering method that would allow us to identify the problem and design possible alternative solutions. On the other hand, their respective class diagram, objects and sequence (important operations within the program) should be attached to show the structure of the program. Finally, each delivery was constantly made (four during the semester), a small 5-minute video should be included in which the new functionalities of the program with respect to that delivery would be shown.

This project allowed us to get closer to a large sample of hard work that depended on good analysis, thorough research, and consistent planning for a good bottom line. In addition, it helps us a lot for our future as Systems Engineers. Working as a team, delegating positions, and managing time is essential to reach a common goal and stand out for doing a good job.

In conclusion, he leaves us teachings that make us better people every day such as: respect, trust, honesty and responsibility. The commitment for each of the participants was good, although it could improve, we are under difficult circumstances where we have more responsibilities than normal and it is difficult to delegate time. Throughout the project, different positions were evidenced that were coupled naturally, such as: Senior Programmer, GUI Programmer, Documentator and Tester. Finally, there were many ups and downs, but that was not a reason to give up and continue standing with what had to be delivered.

Step 6: Reports and specification

Specification of the problem (in terms of input / output) or functional and non functional requiremnts:

Functional requirements:

Name:	R. #1. Load data
Summary:	The program must be able to load the data from the csv file that was chosen for this project. It is located inside the project´s folder in the following path (../../data/Dataset.csv). The data in the dataset are: “AGE”,”JOB”,”MARITAL”,”EDUCATION”, “DEBT”, “BALANCE”, “HOUSING”, “LOAN”, “DEPOSIT” This data will be manipulated by the program during its execution.
In:	csv file
Out:	The data is loaded in the program.

Name:	R. #2. Display data
Summary:	The program must be able to display on the screen the loaded data from the csv in a table using a DataGridView component. The label of the columns represent the attributes and each row represents a record of the table. The data are: “AGE”,”JOB”,”MARITAL”,”EDUCATION”, “DEBT”, “BALANCE”, “HOUSING”, “LOAN”, “DEPOSIT”
In:	The loaded data from the csv file.
Out:	A table with the data of the loaded file.

Name:	R. #3. Filter data
Summary:	The program must be able to filter the data of the table that is displayed on the screen based on a desired attribute (column of the table). The attributes from which the user can choose to filter the table are “AGE”,”JOB”,”MARITAL”,”EDUCATION”, “DEBT”, “BALANCE”, “HOUSING”, “LOAN”, “DEPOSIT”. This option will be displayed using a ComboBox component.

In:	the desired attribute.
Out:	a filtered table will be displayed on the screen in real time.

Name:	R. #4. Show charts
Summary:	The program must be display 5 charts that represent some variables of the dataset. The program must display a bar chart for AGE, JOB and MARITAL. Th program must display a Circular chart for DEBT and HOUSING.
In:	<None>
Out:	5 charts that represent the behavior of a variable.

Name:	R. #5. Classify variable
Summary:	The program must be able to classify a variable using a decision tree. For this particular case the program will classify the clients of the bank which are represented by each of the records from the data table that contains the loaded information. The classes of this problem are yes/no. “Yes” if the client will acquire the subscription to the term deposit and “No” if not.
In:	a record of the table.
Out:	the class of the variable

Name:	R. #6. Train a tree
Summary:	The program must be able to train a tree to using when we should take the predition of some register
In:	Dataset
Out:	Tree train successful

Name:	R. #7. Select the tree

Summary:	The program must be able to select the tree that we would like to use to give prediction about register
In:	<None>
Out:	Own tree
	Library tree

Non Functional requirements:

Name:	NFR. #1. Dataset
Summary:	The program must only read the selected dataset in order to run properly.
In:	
Out:	

Name:	NFR. #2. Programming language
Summary:	The program must be written in the C# language.
In:	
Out:	

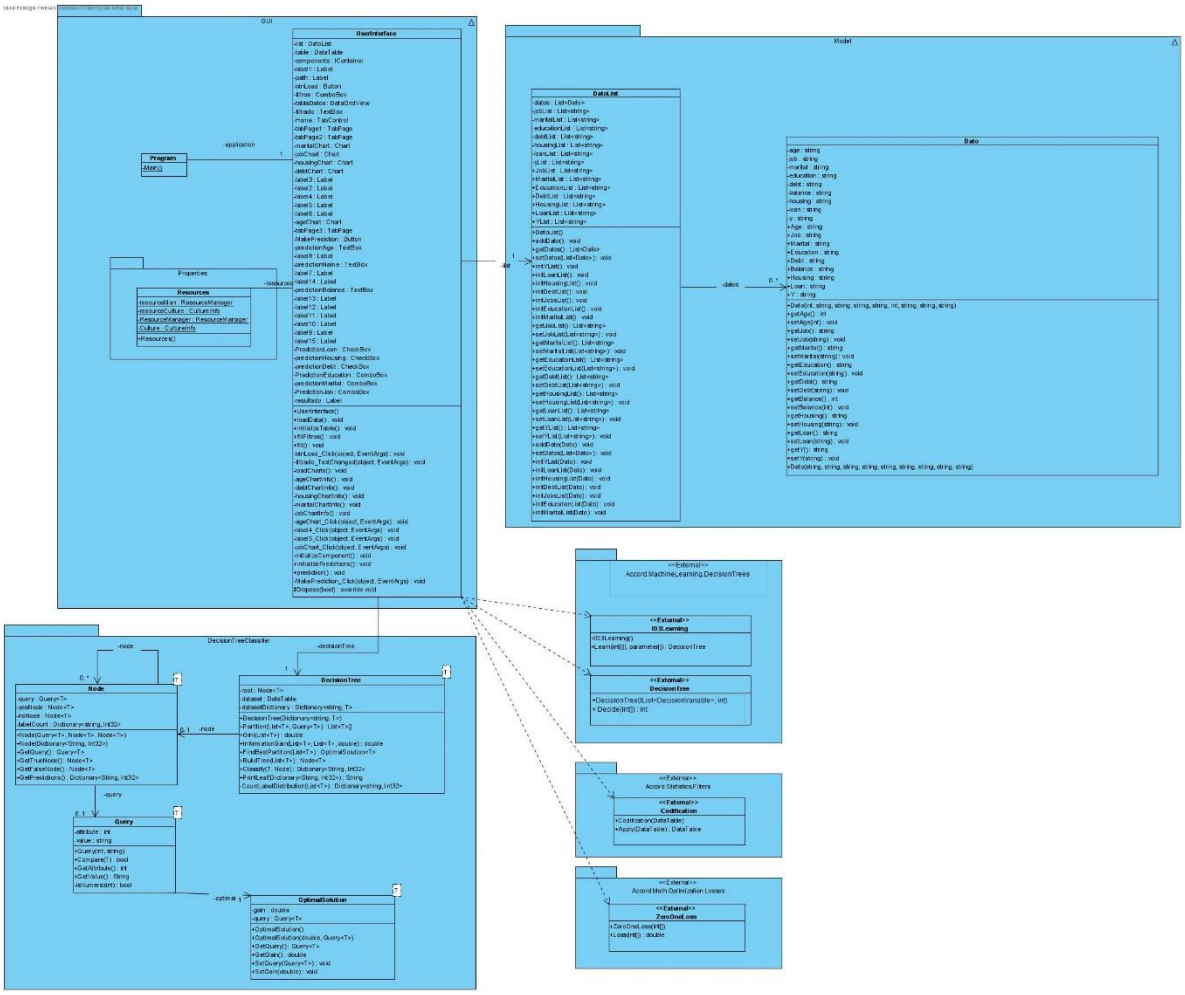
Name:	NFR. #3. Framework
Summary:	The program must be developed using the .NET framework.
In:	
Out:	

Name:	NFR. #4. Tree implementation
Summary:	The decision tree must be implemented by us, not by using an external library.
In:	
Out:	

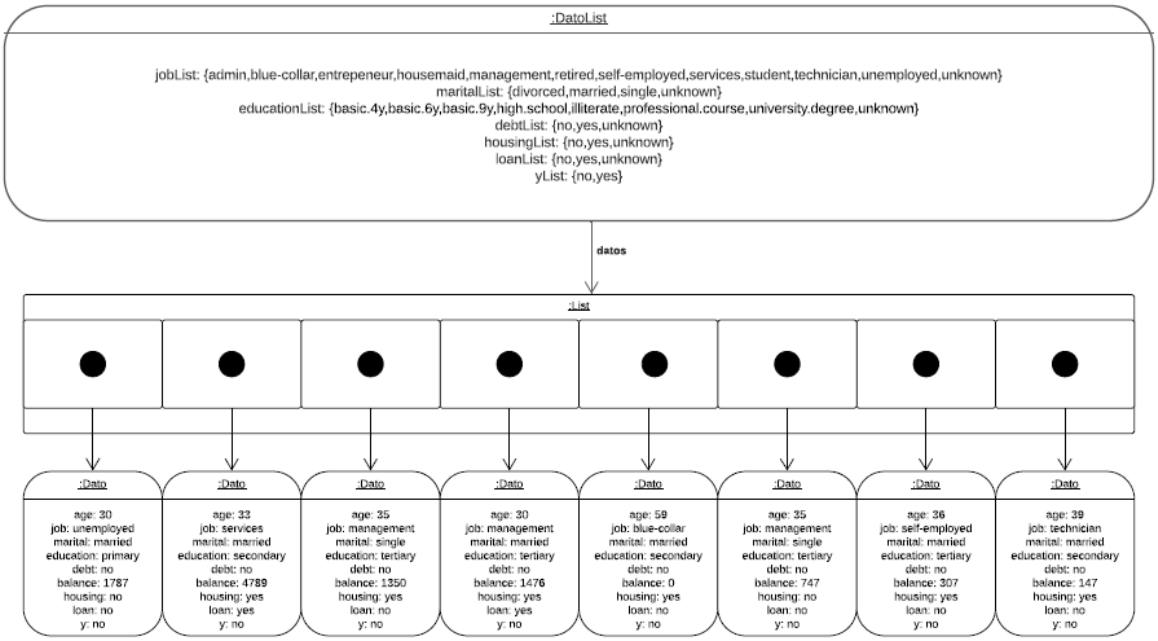
Name:	NFR. #5. Own implementation of tree

Summary:	The decision tree must be implemented by us, not by using an external library.
In:	
Out:	

Class diagram:



Object Diagram:



Bibliography

Chen, J. (16 de 03 de 2020). *www.investopedia.com*. Obtenido de https://www.investopedia.com/terms/t/termdeposit.asp

Geeks for geeks. (2019, 04 17). *www.geeksforgeeks.org*. Retrieved from https://www.geeksforgeeks.org/decision-tree/